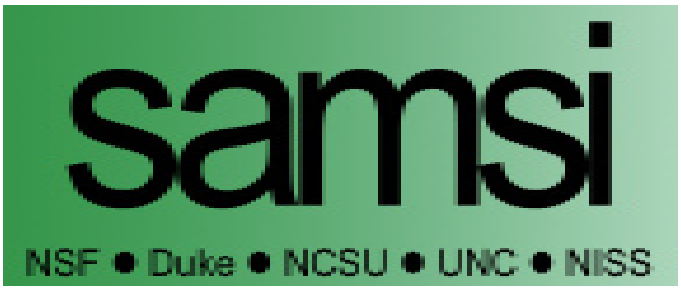


Sieving Genomic Features by Signal Separation and Interference Sparsification

Enrico Capobianco

Technical Report #2007-3
February 26, 2007

This material was based upon work supported by the National Science Foundation under Agreement No. DMS-0112069. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science

Statistical and Applied Mathematical Sciences Institute
PO Box 14006
Research Triangle Park, NC 27709-4006
www.samsi.info

Sieving genomic features by signal separation and interference sparsification

Enrico Capobianco
CRS4 Bioinformatics Laboratory
Tech Park of Sardinia, Loc. Piscinamanna, Ed. 3
09010 Pula (Cagliari), Sardinia - Italy

February 25, 2007

Abstract

Microarrays measure the abundance of thousands of mRNA targets simultaneously, but it is notorious that they deliver limited samples and noisy high-dimensional values for the gene expression changes. This is usually due to the experimental conditions, whose constraints affect the application of artificial learning procedures and require careful adaptation. For instance, when the goal is performing reverse engineering, which consists in reconstructing from the observed measurements the gene-gene interaction network through the identification of the structure and strength of its connections, the inverse problem that has to be solved requires relevant computational efforts in terms of model selection (from statistical learning, information theory, machine learning, signal processing, etc.), regularization (how to deal with under-determinacy from high dimensionality?) and algorithm development (feasible and efficient N-body computation). From the standpoint of building accurate statistics for a reliable inference, the issue of dealing with extremely finite-sized data is crucial. To this end, the fact that experimental replicates exist cannot be underestimated, and indeed can be usefully considered, as shown here. In particular, feature selection performed with projective methods is studied, and it is adapted to exploit the information content of replicates by implementing a simple heuristic sieving idea. Despite its roughness, this strategy offers the merits of inherently questioning the appropriateness of averaging information among replicates, which is a common practice. Then, it looks into the problem of passing from the observed high-dimensionality to the (approximate) intrinsic dimensionality, and this in relation to both the existing sparsity constraints typical of biological networks and the sensitivity of the feature learner. The goal is to stabilize the results with respect to the inherent variability of gene expression data. Correspondingly, there will be more chances of verifying the role that replicates may play for elucidating the complex genomic regulatory maps.
Keywords: *Dimensionality; Genomic Signal Deconvolution; Independent Component Analysis; Sparse Learning; Gene Feature Interference.*

1 Introduction

Genomic signal processing has gained a quite relevant impact in systems biology due to the design and implementation of methods that explore the structural properties of complex biological systems, infer their regulative dynamics, predict the interactions among variables. For instance, when the goal is extracting a map of gene-gene interactions from experimental measurements, one may start from the analysis of generated expression signals, continue with their grouping by pattern similarity, and then apply inference procedures about the latent regulation patterns underlying the observed complex dynamics.

Microarrays represent formidable informative sources in genomics, whose structures require innovative solutions (data analysis and methods fusion, model calibration) and integration between quantitative and qualitative features (merging of biological knowledge and multiplexing of multi-omic heterogeneous information).

From a model building standpoint, many sources of variation have to be accounted at the biological, technical, experimental, and statistical design level. Thus, an ideal strategy would be a flexible one that aims to regularize the inference problem in order to properly address its typically high-dimensional and poorly sampled variables.

Apart from removing systematic errors, it is key to recognize by ad hoc techniques the spurious correlations and the collinear behaviors among genes. The effect is to increase the immunization benefits of the extrapolated dynamics with respect to redundancy and artifacts in the observed data. Also for this reason it is normally accepted to apply some kind of dimensionality reduction to minimize the impact of noise and extract features from the most informative dimensions.

The methodological work we propose aims to investigate the potential of calibrating feature selection for the gene network dynamics that are experimentally observed in the *Escherichia Coli* organism. In particular, we monitor the activation of its *SOS* pathway in response to induced (via time course perturbations) DNA damage. Our focus is on isolating a certain number of gene-gene interactions whose dynamics are the target of the perturbation experiments. Then, especially the impact of noise and signal interference[1] is examined.

Since three experimental replicates are available, we attempt to accurately measure the influence that they have on the feature selection performance of dimensionality reduction algorithms, on the deconvolution of the signal features, and on the sparsification of the interferences of various nature.

The paper describes in Section 2 data dimensionality aspects and how to employ feasible global and local reduction techniques, with a particular emphasis on learning through Blind Source Separation (BSS) approaches. Section 3 deals with microarray-based data mining and processing by Independent Component Analysis (ICA), while Section 4 refers to general signal deconvolution aspects. Section 5 reports the outcomes of our computational algorithms and numerical experiments. Then, Section 6 is for the final remarks.

[1]Interference in signal processing is usually seen as a mix of unknown and time-varying cross-talk and noisy phenomena observed simultaneously, and particularly at high frequencies.

2 Dimensionality and Subspace Analysis

High-dimensional data are studied in several research domains and in general deliver variable degrees of noise and structure convolution. Dimensionality reduction (via feature selection and extraction) and space decomposition (via data partitioning or clustering) methods can achieve denoising and uncover linear or non-linear correlation (often monitored thorough comparisons between Pearson correlation and mutual information).

Many studies have been generated along these lines, and in general, the informative data bulk lies in a much lower dimensional manifold compared to the initial one, which usually needs to be identified and estimated through unsupervised learning methods. Two aspects are also worth to be mentioned: transiency of dynamics, and equilibrium/non-equilibrium aspects.

In the language of dynamical systems, there are phenomena that exhibit transient behavior, followed by an asymptotic motion lying on an attracting set, i.e. a subset of the phase space contained is some finite dimensional manifold. Establishing the invariants of such systems is key, but this step often requires the usage of local adaptive bases, such that one can simply approximate an embedding by a set of local structures.

This first aspect suggests that the goal is encoding the information bulk by a coordinate system that is optimal for some criteria. For instance, finding a minimal projection error of the orbits on the new coordinates while preserving locally the geometries in the low-dimensional space.

Furthermore, global equilibrium structures might just be an ideal way of looking at a complex system or network, which holds only at a particular (steady) state of the system. Conversely, it is more likely that the dynamics observed at a certain discretized grid of time points occur under non-equilibrium conditions, and therefore require a more integrated analysis of the interactions between separated or overlapping local structures.

While assuming smoothness of the data manifold is usually very useful (it allows a sort of complementarity between topological and structural properties regardless we look at them as subsets of bigger spaces or more abstract spaces), one may first consider the problem of finding efficient embedding algorithms that achieve low dimensionality and small distortion.

Thus, a D-dimensional manifold M is embedded in an N-dimensional space, with $D << N$, and represented as a function $\xi : R^D \rightarrow R^N$. A collection of data points $\{X_i, i = 1, \ldots, n\} \in R^N$ is available from a measurement device and can be considered an outcome of noisy sampling from the manifold. From these data, a latent variable model $X_i = \xi(f_i) + \epsilon_i$ can be built, with the aim of estimating the unknown low-dimensional data projections or features f_i.

In statistical terms, the manifold learning problem requires to reconstruct the unknown ξ mapping from the data, which might be done by using non-parametric techniques such as kernel, splines, wavelets. The embedding problem can also be generalized so as to account for a certain distortion with which the input space-manifold mapping may occur. Consider a measurable object $\phi \in \Phi$, its Euclidean norm l_2, and a finite metric space (ϕ, d), with $d(\phi_1, \phi_2) = || \phi_1 - \phi_2 ||$; for a certain factor $\eta \geq 1$, call ρ an example of $\eta-$distorted embedding of ϕ into ψ (to be considered a normed space) such that (ϕ_1, ϕ_2) satisfy the relationship $d(\phi_1, \phi_2) \geq || \rho(\phi_1) - \rho(\phi_2) || \geq \eta^{-1} d(\phi_1, \phi_2)$.

When the chosen norm is linked to the smoothness of the approximation object, it is likely

that a global rather than a local dimensionality reduction technique fails to detect truly useful information (outlying directions, clusters, etc.) in a non-linear manifold, particularly at its curvature points. Consequently, the approximation error (or the distortion degree) should correspond to a truncation error obtained from a global expansion of the object of interest (i.e. where the higher order terms have been discarded). Given a suitable approximation space Φ and an orthonormal basis $\{\theta_i\}_{i=1}^{\infty}$, one can define the error from projecting $o(\phi, t)$ (i.e. the orbit of ϕ, with $t \in [0, T]$) into the first k basis elements as:

$$e^k(\phi, t) = o(\phi, t) - \sum_{i=1}^{k} < o(\phi, t), \theta_i > \theta_i \tag{1}$$

The goals of finding an optimal basis (among all the possible bases) and the best possible ϕ are obtained by minimizing:

$$e^k(\phi) = lim_{T\to\infty} T^{-1} \int_0^T || \ e^k(\phi, t) \ ||^2 \ dt \tag{2}$$

It is well-known that under ergodicity[2] and compactness, a basis of eigenvectors satisfies the problem. This basis is global, so to speak, but it can be further refined to be localized so as to approximate the dynamics specifically in some regions of the invariant manifold. Thus, one might try to characterize highly non-linear manifolds in a locally linear way and with a good overall approximation accuracy, by building for instance probabilistic models that assume (through some priors) the data to be sampled with a certain distribution from different regions of the manifold[3].

The only informative source about the measurable system's dynamics are usually the observed data X, obtained from a finite temporal interval (which corresponds to examining a curve in a finite dimensional space). These data points represent the phase space for a dynamical system; define them by Y_j, with $j \in J$ an index set standing for either a finite set $\{1, \ldots, N\}$, or a certain (time, space) interval $[0, T]$. In many applications, a combinations of these two sets is the best way to represent the available (experimental) data, i.e. $J = [0, T] \times \{1, \ldots, N\}$, which might thus be seen as a collection of trajectories sampled at uniformly discretized points in $[0, T]$.

Genomic data can also be represented this way. One problem is to eliminate the possible redundancies they might be present. Therefore, one needs to look at the genome profiles (or the trajectories computed at each condition), where some but not all genes might be affected by external perturbations or cascading effects. Then, the gene patterns should be monitored, and since they come from repeated measurements (samples) taken at varying conditions, the problem is to distinguish between random fluctuations and temporal correlation dynamics.

[2]The long term temporal average along a single orbit is equal almost everywhere to the average calculated over all the initial conditions.

[3]When each region has Gaussian local coordinates, in this case the model becomes a well-known Mixture of Gaussians.

2.1 Global Decomposition Approach

The dimensionality problem has been already considered in genomics, and tackled by a classical approach such as Singular Value Decomposition (SVD, [1, 19]), likewise, when looking at the data covariance, through Principal Component Analysis (PCA, [25]. One basically seeks a subspace $r \in R^N$ such that given the data Y and the orthogonal projection operator P_r, a total square distance is minimized:

$$|| Y - P_r ||^2 = \sum_{j=1}^{N} \int_0^T || Y_j(t) - P_r[Y_j(t)] ||^2 \, dt \tag{3}$$

Given the corresponding correlation matrix:

$$Cor_Y = \sum_{i=1}^{N} \int_0^T Y_j(t) Y_j^{'}(t) dt \tag{4}$$

and its ordered eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N \geq 0$, then the minimum value of $|| Y - P_r ||^2$ over all possible $k(<< N)$ dimensional subspaces r is given by $\sum_{j=k+1}^{N} \lambda_j$, and the minimizing r is the invariant subspace corresponding to the eigenvalues $\lambda_1, \ldots, \lambda_k$.

Correspondingly, feature selectors based on the SVD pre-processing step induce a decomposition $Y = U \Lambda V^{'} = \sum_{i=1}^{N} \lambda_i u_i v_i^{'}$, where $Y \in R^{N \times M}$, $U \in R^{N \times N}$ is an orthogonal matrix of left singular vectors $u_i \in R^N$, i.e. the eigenvectors of $YY^{'}$, $V \in R^{M \times M}$ is an orthogonal matrix of right singular vectors $v_i \in R^M$, i.e. the eigenvectors of $Y^{'}Y$, and $\Lambda \in R^{N \times M}$ is a diagonal matrix of singular values, ordered such that $\lambda_1 \geq \ldots \geq \lambda_k \geq 0$, with $k = min(N, M)$.

With linear independent signals, all the structures are needed for a perfect reconstruction, as each one carries unique information; vice versa, under some degree of linear dependence, the full reconstruction power can be given by just a few structures. Equivalently, a given statistical ensemble can be represented by a minimal number of modes reflecting the number of degrees of freedom in the systems. As most of the variability is captured by the first subspace, and the residual fluctuations are accounted by the other subspaces, the role of noise in the system is substantially diminished by neglecting part of the latter.

In a genomic setting we have a gene matrix G, with $g_{ij} \in G, i \in \{1 \ldots N\}, j \in J = \{0 \ldots T\}$, with the rows representing time course signals (i.e. gene temporal patterns in the high-dimensional gene space), and the columns representing genomes at a given time (i.e. genomic profiles in the condition space). We can then compute the eigensystem with eigenvalues (i.e. the energies of the modes) and eigenvectors (say γ, determined by maximizing the energy in each mode), and last keep a few dominant eigenvalues $1 \ldots k$ to get a compact representation of the gene matrix, $\hat{G} = \sum_{i=1}^{k} f_i \gamma_i^{'}$.

This results in a generalized coordinate system defined by the eigenfunctions of the gene matrix and under a mean squared error optimality principle, such that $[G - \sum_{i=1}^{k} f_i \gamma_i^{'}]^2 = min$. In particular, the first mode contains the largest proportion of kinetic energy of the signal, while the other modes contain decreasing energy and all together describe the important characteristic features of the system. The described setting offers both statistical and approximation-based interpretations.

With Gaussianity, and orthogonality of the decomposition structures, one finds a decor-

related system, i.e. one where the linear dependence carried by the first two distributional moments has been removed. Thus, $Cor_G = Cor(g_i, g_j) = Cor_{ij} = \lambda_i \delta_{ij}$, for δ_{ij} the Kronecker delta identically equal to 0 for $i \neq j$, and to 1 for $i = j$. Under non-Gaussian probabilistic hypotheses, the decorrelation property holds only approximately, and one can miss important structure in the form of higher order dependence, with a loss of information $L_{ij} = Cor_{ij} - \lambda_i \delta_{ij}$ as a consequence.

2.2 Localized approach via Blind Source Separation

In statistical terms, assume the presence of parameters $\theta \in \Theta \in R^D$ and the existence of an embedding space that via a smooth mapping ρ allow the transformation $\rho(\theta) = X$ deliver a data manifold which agrees with changes in the parameter values. This way, one can address the ill-posed recovery problem of both the parameters θ and the mapping ρ.

To this end, and by following the projection pursuit idea[4] [17, 22], BSS and ICA [8, 10, 14] have been shown to represent possible improvements over the standard PCA. As the main goal is to explain most of the gene variability by only a small number of latent variables, instead of orthogonal and uncorrelated quantities this time the target is the statistical independence among components, which is searched through the maximally non-Gaussian components under a constraint of constant variance. Naturally enough, the dimensionality of the latent space becomes the manifold intrinsic dimensionality of interest.

As far as concerns relevant literature review, some good reasons for applying these ideas and the related methods in genomics have been emphasized in recently published work on microarray data [27, 26, 13], gene feature detection [2, 7] and classification [20, 21].

The main interest here refers to feature selection and extraction; since the projection of gene profiles over a particular template expression pattern delivers the corresponding feature set, say $f \in \mathcal{F}$, the latter quantifies how each gene correlates (positively or negatively) with the chosen template and with variable degree depending on the relationships between the examined genes and the involved biological processes or pathways.

2.3 The Model

The simplest possible linear ICA model is here considered, which represents a mixture of signals as $G = AS + \epsilon$. Given an $N-$dimensional genomic vector and an available sample size equal to the number of conditions, the independent non-Gaussian sources (or components) S have the same genomic dimensionality of the mixture, but are assumed in a number M bounded by the rank of the gene matrix. Last, the linear mixing occurs through the mixing matrix A, and both S and A must be estimated.

Estimating the parameters in the mixing mechanism and recovering the hidden sources imply a certain degree of indeterminacy. In particular, the mixing matrix includes the relative strengths or weights attached to each component, and indicates the type of connectivity map that might exist between the main biological regulators of the gene expression patterns.

Several computational algorithms can be efficiently applied, such as the Joint Approximate Diagonalization of Eigenmatrices (for Real signals) or *JadeR* [10] and the *fastICA*

[4]The projections with densities departing from normality represent the most informative directions to be explored by projective methods.

[23, 24], probably the most popular ones. These algorithms have been adopted here too; while the former offers a better control of the sequence of operations done by ICA, the latter is particularly useful for doing data pre-processing via principal components, and for performing numerical optimization under different conditions.

Theoretical work suggests that if the number of observed mixtures is greater or equal than the number of sources, then it is possible to separate statistically independent sources provided that at most one of them is Gaussian ([8, 24]). Once the mixing matrix is estimated, the components can be obtained by the either the inverse or the pseudo-inverse matrix[5].

Thus, one obtains the estimates of the separating or de-mixing matrix B via $pinv(B)$, and then finds $Y = BX$ that approximate S. Under conditions of perfect separation, the equality $Y = BAS = S$ makes sense, but real solutions hold only approximately, up to permutation P and scaling D matrices, i.e. $Y = DPS$.

The key of the interference problem we aim to address relies on the interpretation of the genomic regions where a complete separation cannot be achieved, because of erratic algorithmic behavior, presence of noise signature, or just biologically motivated reasons.

2.4 Why ICA instead of PCA?

Among several type of projective methods, the classical PCA approach defines a metric that is intrinsic to the data, and determined by their second order statistics. The approach of ICA, instead, relies on higher order statistics and switches from a global to a local metric, thus enabling the exploration of the most outlying informative contents.

One difference between ICA and PCA is that in the former case a nonlinear step is undertaken when the space of solution is numerically explored; the optimization step is pursued via information theory or statistical criteria, such as either entropic or likelihood functionals, respectively, which are also addressed as "contrasts". Naturally enough, ICA learning brings advantages and limitations: while the search in general can only lead to local optima in feature space, there might be a better chance to capture information compared to simple orthogonal decompositions [9].

In genomics, much of the relevance assigned to the dependence structure of the data comes from the gene co-expression patterns from which one would like to infer co-regulation between genes. In particular, regulation influences behind the observed gene dynamics can be more or less independent on each other, because referred to either interconnected or unrelated biological processes.

Furthermore, from the spectrum of eigenvalues obtained from PCA or ICA, the selection of the significant eigenvalues leads to discarding those of similar small magnitude. In particular, we would expect that a low-pass subspace associated to the top eigenvalues would concentrate information about the activity of major regulators underlying genes, naturally linked to an amplitude-related signal space. The band-pass subspace would represent a mix of dynamics related to several biological activity (proteins, metabolites, etc.) occurring at finer time scales, and the high-pass subspace can usually represent the noise effects intertwined with the highest resolution dynamics.

[5]When there is no unique inverse, owing to the existence of an infinite number of independent components which are solutions of the linear problem, the latter is under-determined. This case requires constraining the sources or transforming the model in order to get an approximate solution.

3 ICA Mining of Microarray Data

3.1 Sparsity

In dynamical systems one often seeks to estimate the norm of the trajectory of a linear time invariant system in a finite interval and in response to a forcing input u. Consider the differential system:

$$\dot{X} = AX + u, \quad X \in R^N, u \in R^N, X_0 \in [0, T] \tag{5}$$

whose solution is:

$$X(t) = \int_0^t e^{A(t-\tau)} u(\tau) d\tau + e^{At} X_0 \tag{6}$$

In discrete dynamics form the system becomes:

$$X = M_t^A(u) + N_t^A(X_0) \tag{7}$$

The norm estimation task is easier when less redundancy is involved, or equivalently with a sparser matrix structure to be learned.

Exploiting the natural degree of sparsity is crucial in gene regulatory dynamics, where the gene connectivity matrix measured under specific conditions often discards any redundancy from typically dense and highly structured networks. Several attempts have been made to describe gene networks with only limited interactions, especially when it is possible to consider prior structural or qualitative information about the biological system under study and its topology.

In highly adaptive organisms most of the genes have few regulators and in turn regulate few genes; thus, inferring biological co-regulation requires some form of regularization that reduces the dimension of the search space, and still locally detects the key gene relationships. Due to undersampling, noise and high dimensionality, statistical inference that attempts to explain the gene dynamics often result in regression models severely affected by collinearity among the predictive variables.

Consider G_t as the $N-$vector (state) gene expression level representation, with $G_t = [g_{1t} \dots g_{kt} \dots g_{Nt}]$, and T_G as the transition matrix. The gene dynamics are assumed from a continuous time dynamical system of differential equations $\dot{G} = T_G G$, and are thus studied through a discretized system where the goal is to find a stable T_G, i.e. $\bar{T}_G$, which would allow to build recursive estimation of the state vector G and approximate the solution trajectory values according to a difference equation like $G_t = \bar{T} G_{t-1}$. The target matrix represents the strengths of the influence between all genes within the minimal interval allowed by the available time resolution.

It would then be possible to infer the gene regulatory network by applying statistical optimization criteria; since the model is described by linear and time-invariant dynamics, the explained gene interactions could be justified by a linearization of the underlying complex non-linear system around a stationary point, while a useful extension would be to model

unknown factors through the insertion of hidden variables S_t and noise ϵ_t, thus leading to a latent variable stochastic model such as $G_t \approx AS_t + \epsilon_t$.

Naturally enough, S_t may represent effects on G_t that cannot be measured in a gene expression profiling experiment, such as the level of regulatory proteins, or biological influences due to unobservable metabolites, or various mRNA and protein degradations. Thus, one might hope to recover part of the complex dynamics which are lost at the first order degree of approximation.

Furthermore, the dynamic systems's norm estimation problem now restricts the analysis to just the latent information sources S and their mixing mechanism A[6]. An advantage of ICA is its flexible structure, just a simple *signal + noise* model free from specified probabilistic assumptions on the underlying distributions[7].

3.2 Data and Application Goals

For the present study, the *Escherichia coli* bacterium is investigated and its DNA repair activity monitored through the *SOS* response system. The genomic measurements are generated by time-course perturbation experiments described as follows.

At time zero, when the gene expression patterns are in equilibrium, a drug called (*Norfloxacin*) is added to the cell cultures. As the cell's environment changes, the gene expression patterns are monitored to detect how they adapt and change, until a new cell's equilibrium is achieved (steady state conditions) after some time (approximately one hour).

Between the two equilibrium points various intermediate cell's states are measured according to a pre-determined sampling process that occurs at regular discrete time intervals (every twelve minutes) and delivers six measurements (two extra conditions refer to no-drug states, for control). Three replicates are performed at each condition.

The microarray data are then processed by the *Affy MAS 5.0* algorithm, normalized with respect to expression values recorded at the start time of the experiments, and transformed by logarithms in order to further control the extreme variations.

We underline the relevance of replicates for biological validation of the results. In particular, it is common before applying any feature learner to average the gene expression measurements observed from various experimental replicates.

However, discarding information about this kind of variability can be detrimental for establishing reliable levels of confidence in the learning model, whose assessment would benefit from such neglected information.

Thus, one would need to know how useful is to have a certain number of replicates and coherent degree of replicability with a clear scope in mind, that the inherent variability of gene values might be accounted for by feature selectors, together with the extent to which consistent and reliable findings can be found across replicates.

We address a problem of redundancy vs sparsity bounds in the system under study; the hypothesis to verify is whether a full (with all the replicates included) or a restricted (just

[6]Identifying this model and estimating the system's parameters is possible by Bayesian learning methods or classical regression models; however, both these techniques involve quite intensive computations, and still require regularization in order to deal with the ill-posed problem, thus learning the network sparsity structure by either priors or dimensionality reduction.

[7]In statistical terms, ICA can also be considered a semiparametric inference technique [11, 12].

one set of measurements from averaged data) model results in a decreased risk of missing some relevant gene relationships.

3.3 Gene Selection via Thresholding

By using ICA we aim to address gene selection from microarray data. The identification of gene groups is achieved by extracting signals from noisy data by thresholding and construction of statistical confidence intervals. The sets of genes endowed with outlying informative content (i.e. more or less differentially expressed compared to an average value) are isolated in each estimated component. Table 1 illustrates the size of the feature sets identified at different intervals by the application of ICA to each replicate.

The outlier isolation process [6] performs gene selection over the estimated components via fixed intervals; we take a scalar q times the standard deviation from the mean of the pivotal estimated component, and identify as outliers the genes whose values are outside the interval.

The simple algorithmic steps that are repeated for each estimated component may be given this way:
1. Fix a pivot component j;
2. Calculate an interval at level fixed by q (equal to 2 or 3, and approximately the 95% or 99% confidence levels);
3. Compute the mean and the standard deviation over the pivotal $N-$dimensional genomic profile selected in turn;
4. Consider next component $j+1$, for $j = 1, \ldots, M$.
The corresponding interval[8] is:

$$mq_N^j = mean[s_N^j] \pm q * std[s_N^j]. \tag{8}$$

Thresholding exploits the discriminatory power of each estimated component. The selected genes that are significantly deviating over the gene profile in each estimated component support the rejection of the null hypothesis of no presence of outlying effects within the ICA-based genomic profile.

As the components represent as much as possible statistically independent information, one can hope that up to a certain approximation degree this independence may reflect biological reasons. While genes sharing the same biological processes or belonging to the same pathway are expected to show relatively strong connectivity, and should thus be grouped together, it is also possible that some genes may belong at the same time to different groups, depending on co-regulated dynamics.

4 Signal Deconvolution

Recovering the structure underlying an observed gene network may require signal deconvolution strategies that address an ill-posed problem and usually offer with non-unique solution.

[8]The term mq represents the gene expression mean value calculated as log-ratio.

ICA-2	*Replicate 1*	*Replicate 2*	*Replicate 3*
IC1 - m2	114	104	113
IC1 - m3	14	17	16
IC2 - m2	93	82	96
IC2 - m3	24	26	29
ICA-3	*Replicate 1*	*Replicate 2*	*Replicate 3*
IC1 - m2	113	104	114
IC1 - m3	12	19	17
IC2 - m2	93	82	96
IC2 - m3	24	26	29
IC3 - m2	124	117	134
IC3 - m3	47	40	40
ICA-4	*Replicate 1*	*Replicate 2*	*Replicate 3*
IC1 - m2	122	111	115
IC1 - m3	12	21	16
IC2 - m2	86	83	92
IC2 - m3	22	25	28
IC3 - m2	116	119	134
IC3 - m3	51	39	40
IC4 - m2	125	107	122
IC4 - m3	38	56	52

Table 1: Summary statistics - ICA with 2 (top), 3 and 4 (bottom) components. The size of the estimated components for *m2,m3* intervals and all the replicates (*Replicate 1-3*) is reported.

In order to extract independent sources, ICA leads to the exploration of the least Gaussian one-dimensional projections, or equivalently of the directions provided by $B = pinv(A)$.

A stochastic component comes from the between-replicate variability, which increases the uncertainty amount already determined by the ICA approximation error and the ambiguities from the unknown A and S. One would expect that as a result of examining all the replicates, a better signal-to-noise ratio is achieved, which means a certain decrease in variation for the biological versus the noise component.

In other terms, from an initial coarse feature selection we proceed to a further comparative analysis of feature selectors by shrinking the sub-sets of genes grouped in each identified component to a minimum set representative of all the replicates, and testing the reliability of our learning system. Thus, we explore measures of synthesis that aggregate the observed replicated information and detect significant discrepancies.

To simplify the problem, three cases (seen as variants of typical majority voting schemes) might be considered:

- a gene is selected by a certain component in all the replicates, in which case there is a strong indication to select it;
- a gene is detected by the component in two over three replicates, which thus should be carefully checked;

- a gene is detected in only one replicate and not in the others, such that it becomes a good candidate to be discarded.

Even under a full agreement across replicates, the effect might be only a reduction of the impact of false positives. Conversely, a weak indication for selecting a gene carries always the risk of discarding differentially expressed genes (false negatives).

As we apply a feature selector to all the replicates, and then compare its relative performance, the goals are:

1. To assess the stability of the feature selection process, which says something about the robustness of source separation and information retrieval;

2. To measure both gene variability and noise sensitivity of each replicate;

3. To establish sparsity bounds for interference sparsification purposes.

4.1 Dealing with Interference

The components might be considered connected or disjoint according to the presence of genes that overlap or not. The extent to which this occurs causes a gene partition between components that despite the imposed statistical independence reveals the presence of underlying relationships between the features.

Given a number m of extracted components, and considering pairwise component relationships i and j, a partition $G^i = \cup_i g_i$ is finer than a partition $G^j = \cup_j g_j$ if some or all the genes g_i overlap more than g_j with genes in other components. Therefore, the corresponding component itself is finely or coarsely partitioned, delivering in case of wide overlapping what might be called a substantial feature interference.

In each component, one could find overlaps of genes in the component j with a number of l components, thus forming gene cross-component subsets such that $G^j = g_{1j} \cup g_{2j} \cup \ldots \cup g_{lj}$. As some genes in a component may overlap with different components at the same time, the effect is of increasing the feature interference[9].

A simple description is offered by the *interference degree*, defined as $INT_D = \frac{Card\{G^{j \cap i}\}}{Card\{G^j\}}$, for $i \neq j$ (with $Card$ indicating the cardinality of the set and $\cap$ the intersection), which is the ratio between the genes in the components i and j that overlap relatively to the whole set of genes belonging to the component j itself.

Correspondingly, another description comes from the *interference intensity*, which represents an energy ratio $INT_I = \frac{\sum_{i=1}^{k}[G^{j \cap i}]^2}{\sum_{i=1}^{n}[G^j]^2}$, for $i \neq j$ and $k \leq n$, where the cumulative energy for the interference region between the components j and i is compared to the whole energy in j, thus giving its relative contribution to its global norm.

These two interference measures refer to a sort of weak connectivity map (gene-gene relationships across components detected through co-expression) linked to an extent to be measured to an underlying physical map (induced by direct, indirect and cascade co-regulation

[9]The dependence degree of the components could be measured by Mutual Information, through $MI_G^{ij} = \sum_{i \neq j} I(G^i, G^j)$, and with respect to different i, j components. This measure will be positive, or zero when two components are disjoint (i.e. do not share any gene).

effects). Overall, the gene network sparsity and redundancy degree also depend on these measures.

Some useful criteria [29] may be adopted to check that the components preserve the specificity of their information, and remove the redundant interferences. Component sparsity represents the property that a small percentage of the signal coefficients captures a large percentage of the signal energy.

Thus, by fixing a certain reference sparsity bound in terms of the gene fraction concentrating most of the energy, an expression threshold value $\bar{G}$ could be optimized according to:

$$\bar{G}^* = argmax_{\bar{G}} \frac{\sum_{k:|G(k)|\geq\bar{G}} \mid G(k) \mid^2}{\sum_k \mid G(k) \mid^2} \geq \mu \tag{9}$$

where $k :\mid G(k) \mid \geq \bar{G}$ is the number of genes that relatively to k (the genes in the component) show high differential expression with respect to the threshold $\bar{G}$, and $\mu = \frac{k:|G(k)|\leq\bar{G}}{k}$ a is a given fraction μ of genes with negligible expression values.

Thus, imposing a given sparsity level could be seen as a characterization in terms of the maximum percentage of genes with small expression level that we can discard while maintaining most signal energy. The optimization problem suggests to search for parsimony, i.e. a minimum number of genes that allow the component to carry at least a certain energy, and this has an impact also on the level of interference.

The energy threshold $\bar{G}$ and the fraction level μ modulate the de-noising power and may induce separation of the signal from noise in the interference regions. With a low percentage of genes in a component that are beyond a fixed threshold, one might measure the corresponding contribution to the total energy and verify whether this ratio is too low (the selected genes might be only moderately differential) or high (sparsity allows for many genes to be eliminated).

4.2 Stabilization and Sparsification

In the present context, stabilization[10] aims to diminish the variability and increase the robustness of feature sets by sparsifying the information identified by the parallel feature learners.

In particular, we aim to control this aspect with regard to the ICA performance in balancing between redundancy and sparsity. A reduction of the interference among the components, in particular with regard to the noise and error contributions, can deliver a clearer picture of the underlying gene dependence (i.e. co-regulation) map.

Given a set of parallel feature selectors, averaging there contents is a common way to combine the distributed information and eliminate the strong fluctuations across replicates. Feature selection performed through sieving or nesting involves a sort of filtering of genes so as to achieve sparsification with respect to the irrelevant information.

Each replicate can be regarded as an information channel whose input sources are mixed together; it is likely that redundancy in the system may rise questions about reliability of

[10]According to [5], an unstable learning algorithm is one for which small perturbations in the training set can produce large changes in the derived model.

the replicates, and thus suggest a possible reduction of their number.

In microarray studies, the analysis of the inherent variability (i.e. biologically related) in gene expression data compared to other sources of variability (i.e. noise, measurement error, algorithmic artifacts etc.) is a necessary step to distinguish between primary and secondary effects generated by the experiments, where the latter effects are usually only indirectly induced (cascade dynamics). This concept is next discussed at a deeper level of abstraction, by looking at the intersection of functional subspaces spanned by the computed components.

4.3 Error Deconvolution

Several error components need to be considered in gene expression values obtained from replicated experiments. In standard model building approaches, additive decompositions are usually adopted for dealing with biological and technical variation[11].

By combining the outcomes of feature learners we somehow stick to the orthogonality assumption by smoothing all the distortion effects which can be measured from both endogenous and exogenous interference, where the former links to co-regulation, and the latter is a by-product of experimental and measurement errors, plus approximation errors from ICA and algorithmic artifacts.

By emphasizing model aspects as in [18], consider the following relationship between the original and the estimated sources:

$$\hat{S} =< \hat{S}, S > S + \epsilon \tag{10}$$

with the inner product term (or correlation) $< \hat{S}, S >= \sum_t \hat{S}(t)S(t)$, and the squared norm (energy) of $\hat{S}$ defined as $|| \hat{S} ||^2 =< \hat{S}, \hat{S} >$.

The error term $\epsilon = \sum_k < \hat{S}, z_k > z_k$ is generated by a zero mean and unit variance Gaussian noise process z, and contributes to noise-related interference sources which are orthogonal to both the correlation core and the residual interferences in the system, and typically depends on experimental and measurement errors that can be easily detected in earlier stages of the data mining process[12].

These two quantities, energy and correlation, suggest a distortion measure whose size is proportional to the source estimation accuracy:

$$D = \frac{|| \hat{S} ||^2 - |< \hat{S}, S >|^2}{|< \hat{S}, S >|^2} \tag{11}$$

In Eq. (8), the correlation term $< \hat{S}, S > S$ accounts for linear dependence across all the components. Thus, the estimated source decomposition can be further refined by splitting the correlation into source-dependent (C_S) and interference-dependent (C_I) terms as:

[11] Biological variation results from various forms of heterogeneity that are inherent aspects of the investigated biological factors, and reflect the variation among all the experimental units that are employed. Technical variation depends instead on the quality of the samples and their preparation, labelling, hybridization etc.

[12] For instance, when genomes are trimmed for discarding big outliers (together with unaffected genes).

$$\hat{S} = C_S + C_I + \epsilon = C_S + [C_{I_{en}} * C_{I_{ex}}] + \epsilon \tag{12}$$

where the deconvoluted term $C_{I_{en}} * C_{I_{ex}}$ represents the interference bulk that once untangled allows the residual term to be considered just noise.

In principle, one could apply SVD so as to separate noise from structure in the interference regions; however, the problem is that subspace separation might be difficult with relatively small interference amount. Another possibility [18] would involve either random or orthogonal projections of the sources onto their span, or the one mixed to noise. By calling P_s and P_{sz} such operators, one obtains the modified decompositions (for interference and noise, respectively):

$$C_I = P_s\hat{S} - < \hat{S}, S > S \tag{13}$$

$$\epsilon = P_{sz}\hat{S} - P_s\hat{S} \tag{14}$$

These projections can be described in either decorrelated or correlated contexts. In the former case one would get $P_s\hat{S} = \sum_k < \hat{S}, S_k > S_k$[13]. With noise in the span, then under orthogonality we would have:

$$P_{sz}\hat{S} = P_s\hat{S} + \frac{\sum_k < \hat{S}, z_k > z_k}{|| \ z_k \ ||^2} \tag{15}$$

5 Outcomes

The error deconvolution model can be investigated through heuristics, and by using the replicate data. This way an advantage is avoiding assumptions such as knowledge of the original sources and mixing mechanisms generating the interference, or even the probabilistic hypotheses behind noise.

The results from aggregation involve two main aspects: the performance of the feature selectors in each component, where the focus is to isolate specific biological aspects; the performance of the feature selectors between components, where the focus is to uncover relevant gene co-regulation dynamics from the signal interference[14].

In Figure 1 we show the temporal patterns of the estimated mixing matrix coefficients in the *ICA-4* case (ICA system with four identified components) for each replicate[15]. Notice the change of sign in some components across replicates, due to the indeterminacies, and the different degree of smoothness of the patterns, as from the inherent variability in each replicate.

The diagnostic plots in Figure 2 and Figure 3 describe the distributional properties of the estimated component. For instance, in Figure 2 we show the case of *ICA-3* (ICA system

[13]Otherwise, a least square solution is given by $\sum_i \kappa \frac{<\hat{s}, s_i>}{cov(\hat{s}, \hat{s})}$.

[14]For annotation results, *ecocyc.org* has been queried.

[15]The online freely available JadeR and fastICA MATLAB packages for ICA identification and estimation have been implemented.

with three identified components) with both histograms and boxplots, where the plot for the second replicate are larger for visual effects. In Figure 3 the genomic signal profiles of each extracted components are instead reported[16].

5.1 Feature Selection

Table 2 shows some descriptive statistics. For each ICA system applied to each replicate and allowed to vary with different approximation power (as from the number of identified components, i.e. *ICA-2, ICA-3, ICA-4*), we compute heuristic "sparsity bounds". Roughly speaking, we extract from each replicate a gene set by sieving both the gene groups isolated in each component and the interference regions at the *m2* and *m3* intervals.

Correspondingly, one would first control in each replicate the size and structure of each gene group, either isolated by a component or overlapping across components. Following this *pre-nesting* phase, an intersection step defines the *post-nesting* phase where the size of the gene groups can be conserved or reduced according to the sparsity bounds achieved by a particular replicate (the sparsity driver) and leading to interference sparsification in the remaining replicates.

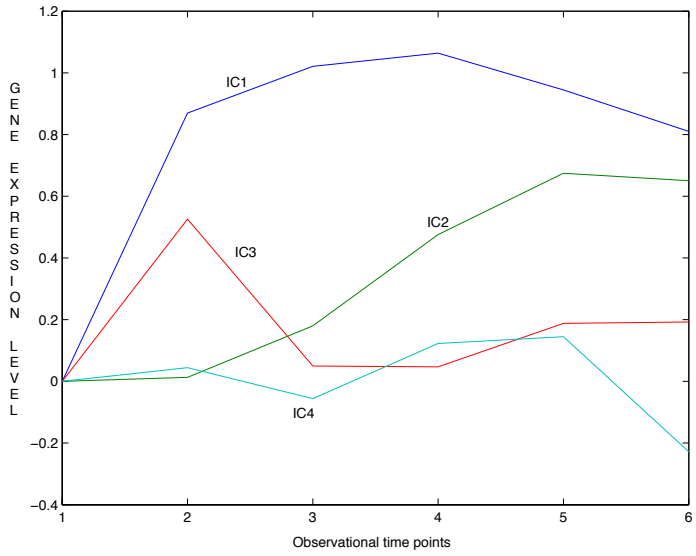
Thus, gene reduction rates can be visualized for each replicate data with respect to the reference (nested) gene sets. The plots reported in Figure 4 and Figure 5 sketch (respectively for *ICA-2*, and *ICA-3* combined with *ICA-4*) the amount of variability in each replicate (from *m2* intervals), and also the big outliers' impact (from *m3* intervals).

Components	SP (*m2*)	SP (*m3*)	RR (*m2*)	RR (*m3*)
ICA-2				
IC1	62	9	$40 \sim 45(\%)$	$35 \sim 47(\%)$
IC2	69	22	$16 \sim 28(\%)$	$10 \sim 15(\%)$
ICA-3				
IC1	56	9	$46 \sim 51(\%)$	$25 \sim 52(\%)$
IC2	40	18	$52 \sim 58(\%)$	$25 \sim 38(\%)$
IC3	40	12	$65 \sim 70(\%)$	$70 \sim 75(\%)$
ICA-4				
IC1	54	8	$51 \sim 56(\%)$	$33 \sim 62(\%)$
IC2	39	18	$53 \sim 58(\%)$	$18 \sim 36(\%)$
IC3	41	14	$65 \sim 69(\%)$	$64 \sim 73(\%)$
IC4	10	2	$91 \sim 92(\%)$	$95 \sim 96(\%)$

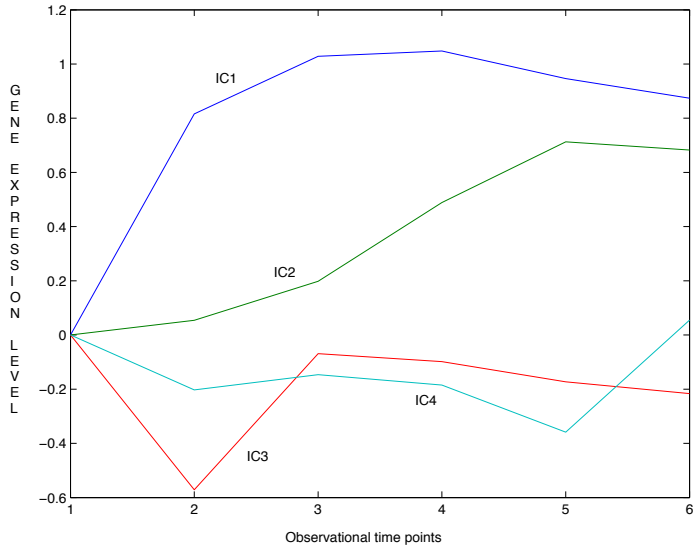
Table 2: Summary statistics - ICA with 2, 3 and 4 components. Sparsity bound (*SP*): number of genes obtained from maximally nested (from all the replicates) component-wide groups. Percentage reduction rate (*RR*): min $\sim$ max values across replicates.

Table 2 describes model selection aspects. From the comparison between the three ICA systems it appears that four components are too many; the corresponding reduction rates

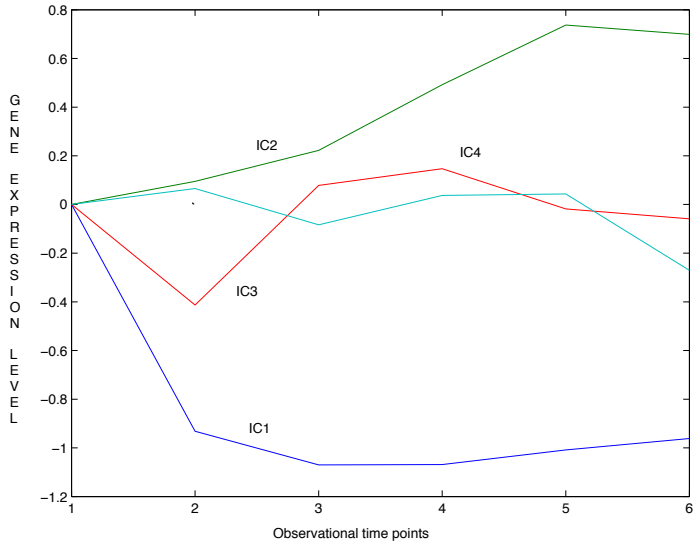
[16]Complete tables of results are available from the author, likewise the entire set of plots produced by the statistical MATLAB 7 toolbox.



A.



B.



C.

Figure 1: Mixing matrices estimated for *ICA-4* in replicate 1 (top), 2 and 3 (bottom).

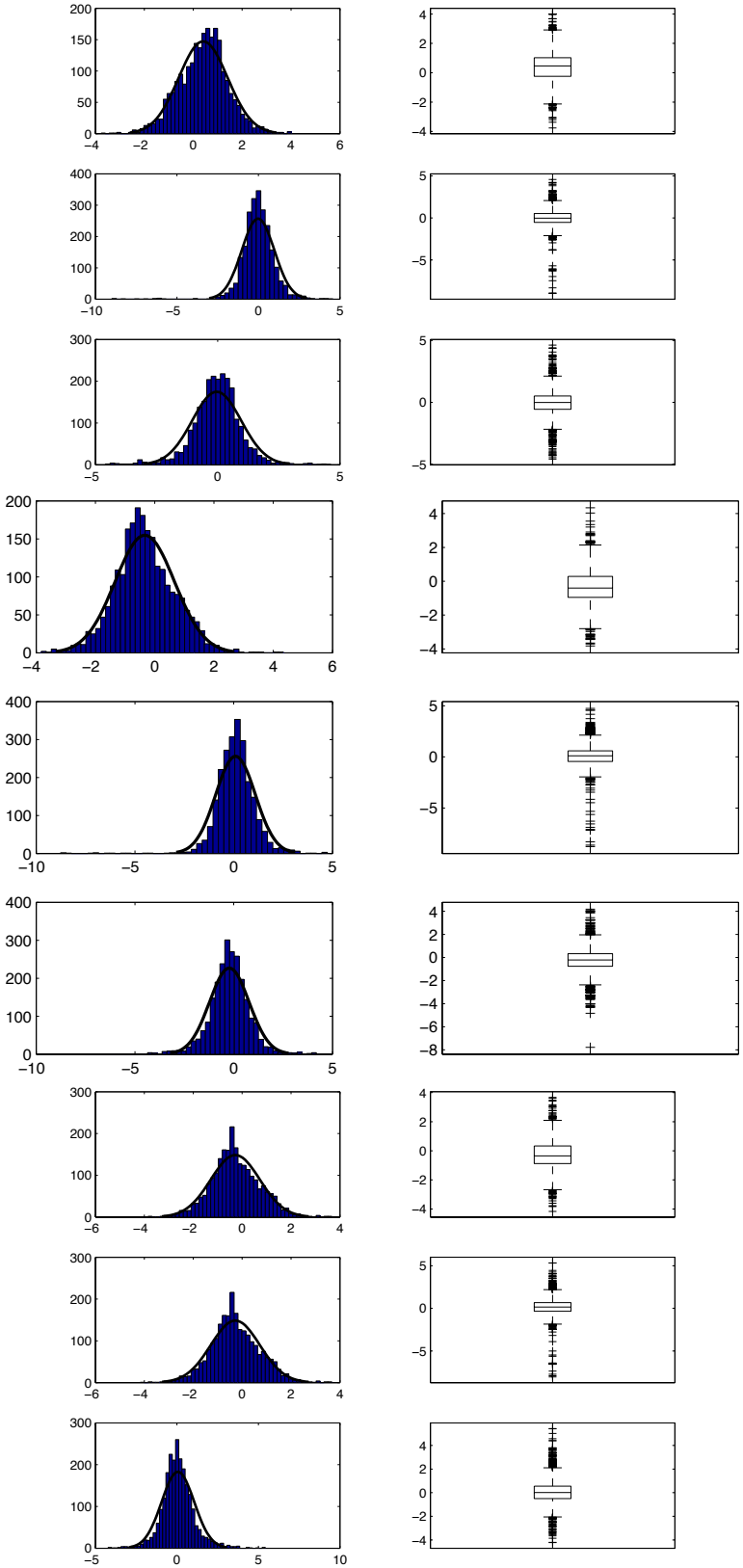


Figure 2: Histogram and box-plots of the *ICA-3* system. *Replicate* 1 (top three rows of plots), *Replicate* 2 (mid three rows of plots, a bit bigger), *Replicate* 3 (bottom three rows of plots). Each subgroup lists IC1 (top) - IC2 (middle) - IC3 (bottom).

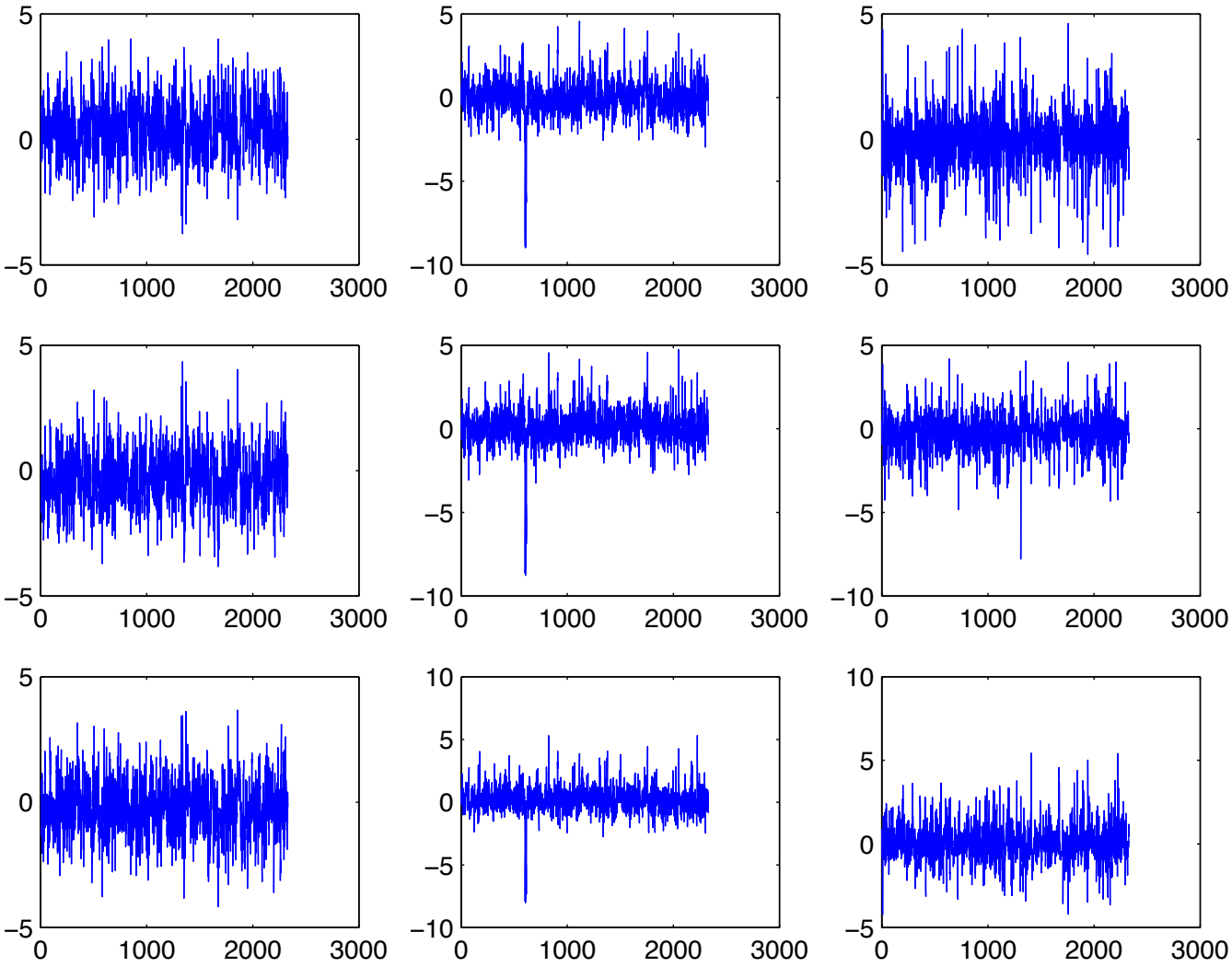


Figure 3: Component patterns in the *ICA-3* system.
Replicate 1 – top-most plots.
Replicate 2 – mid plots.
Replicate 3 – bottom plots.
IC1 (left) - IC2 (center) - IC3 (right).

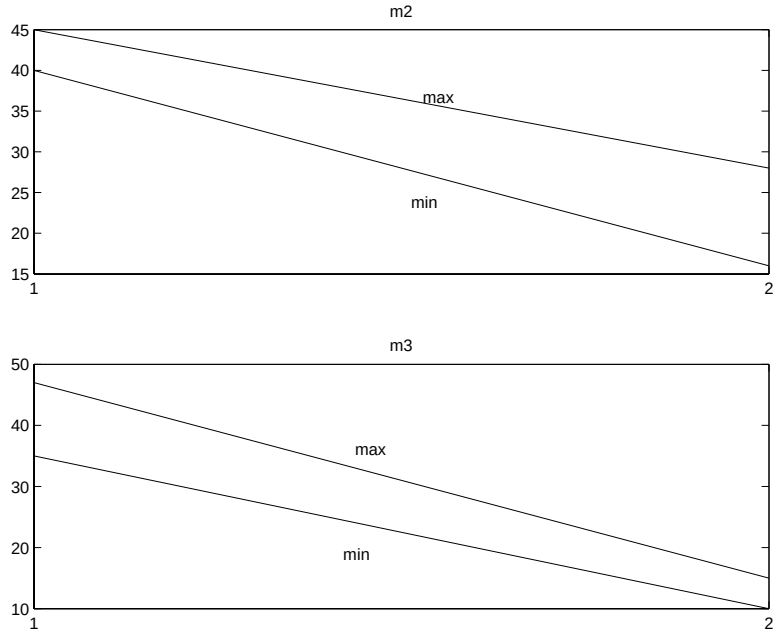


Figure 4: Number of components on the X axis; percentage reduction rates on the Y axis. Rates (min = lower curve, max = upper curve) for the *ICA-2* case. *m2* values (top), *m3* values (bottom).

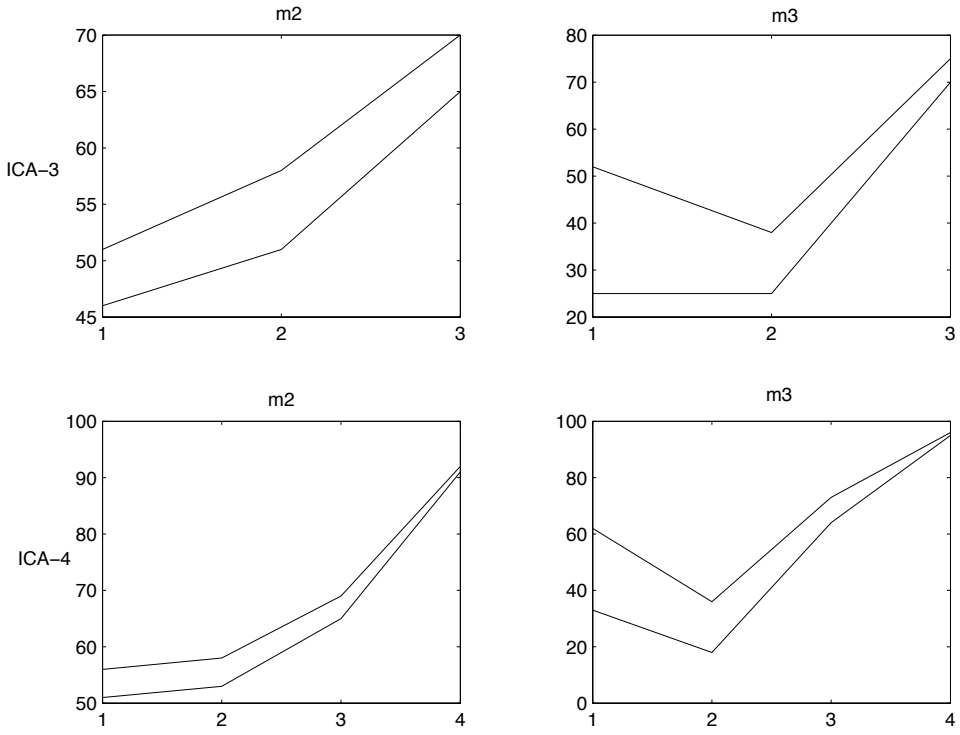


Figure 5: Number of components on the X axis; percentage reduction rates on the Y axis. Rates (min = lower curve, max = upper curve) for the *ICA-3* (top) and the *ICA-4* (bottom) cases. *m2* values (left), *m3* values (right).

are very large for the fourth component, which indicates a big role of noise. The statistics for the first three components of ICA with three and four components are similar, while ICA with only two components is more stabilized.

In qualitative terms, the first component in each system is always the most informative and selects a good sample of *SOS*-involved genes such as *polB* (b0060), *ybfE* (b0685), *uvrB* (b0779), *sulA* (b0958), *dinI* (b1061), *yebG* (b1848), *recN* (b2616), *recX* (b2698), *recA* (b2699), *uvrA* (b4058).

Note that this set appears from each replicate for *ICA-2*, but only with replicate 3 for *ICA-3* (while the other two replicates miss *polB*), while for *ICA-4* the set except *polB* is always selected. The second component in every system is never including genes of the relevant pathway under study, while the third component interferes with the first one in both *ICA-3* and *ICA-4* relatively to *polB*, *recN*, and *recX*.

By looking at Figure 4, we observe that larger percentages of genes are removed from the first component across replicates in the *m2* intervals. The first component, as said, is always capturing most of the variability; thus, one may see that the noise factor has been limited by sparsification.

The gap between the lines indicates the min-max oscillation range in reduction rate, which is larger for the second component; this means that relatively more variability is present across replicates for this component. With *m3* intervals, together with the gene percentage range, also the increased gap for the first component might indicate that a certain outlier effect from some genes is present, particularly when we look at the highest possible (absolute) expression value changes.

In Figure 5, the *m2* intervals (*ICA-3* case) show a quite homogeneous gap across components, which indicates similar variability across replicates, and an increased sparsification with the higher components, probably due to a larger noise contribution. The *m3* intervals present a larger gap in the first component, due to likely outlier effects.

The component sparsification for *m2* intervals (*ICA-4* case) in Figure 5 appears also homogeneous with respect to the presence of noise, with a slightly variable min-max oscillation range among the components, and with patterns in the *m3* intervals that emphasize possible outlier effects in the first component (the one with the largest gap).

Overall, given the size of the *m2* intervals in each component and system, the sparsification effect of nesting through replicates is quite substantial. Now, interference aspects need to be evaluated.

5.2 Component Interference

Gene interference dynamics are examined with the help of ad hoc tables showing the impact of sparsification through the nesting procedure. The structure of the interference regions refers to component doublets computed for each replicate data set[17]. Thus, *pre-nesting* for each doublet indicates all the overlapping genes, while *post-nesting* indicates what is left after sparsification; the genes highlighted in bold font form the nested gene set.

A few other genes are excluded from the final set, but are reported because despite the

[17]For simplicity, we just present results for doublets, but some genes may of course belong to even more components at the same time, such that also triplets, say, could be considered.

exclusion they represent differential genes, either selected in *m3* intervals or appearing in each replicate for one of the two components, but not for the others.

The nested gene set is considered to represent the lower sparsity bound because consists of genes always selected in each replicate by the feature learner. In Table 3 there is the ICA system with only two components, in Table 4 the ICA system with three components, and in Table 5 the ICA system with four components.

The identifying labels instead of the gene names have been reported for clarity. The genes have always been selected by *m2* intervals, thus guaranteeing that the size of the final set is not too small and sensitive to strong outlying effects. One way to adapt them is to consider the genes selected also at the *m3* intervals. However, these entries in the final nested gene set may lead to less sensitivity to false rejection rates.

Replicate 1									
IC1 vs IC2	*b*0809	*b*0810	*b*0811	*b*1085	*b*1109	*b*1137	*b*2656	*b*3285	*b*3418
IC1 vs IC2	–	**b0810**	**b0811**	–	**b1109**	**b1137**	*b*2656	*b*3285	–
Replicate 2									
IC1 vs IC2	*b*0048	*b*0266	*b*0296	*b*0730	*b*0810	*b*0811	*b*1109	*b*1137	*b*1328
	*b*1846	*b*3285							
IC1 vs IC2	–	*b*0266	–	–	**b0810**	**b0811**	**b1109**	**b1137**	–
	–	*b*3285							
Replicate 3									
IC1 vs IC2	*b*0014	*b*0015	*b*0809	*b*0810	*b*0811	*b*1109	*b*1137	*b*3873	
IC1 vs IC2	–	–	–	**b0810**	**b0811**	**b1109**	**b1137**	–	

Table 3: Interference (*pre-nesting* and **post-nesting**) for *ICA-2*. The numbers indicate gene identifiers.

5.2.1 System with two components

The nested set for IC1-IC2 in Table 3 includes some genes; in particular *glnP* (b0810) and *glnH* (b0811) are sub-units of glutamine ABC transporter, respectively membrane component of the ABC transporter and periplasmic glutamine-binding protein; *ndh* (b1109) is related to NADH cupric reductase and NADH dehydrogenase II, and *ymfD* (b1137) is an hypothetical protein.

Clearly enough, the bound is determined by *replicate 3*, while the highest interference level is due to *replicate 2*, which correspondingly becomes the most sparsified component. In particular, *smf* (b3285, an hypothetical protein), which appears in doublets from two replicates, has been discarded. Overall, the detected interference is uninformative, thus nesting has a limited value in this case.

5.2.2 System with three components

In Table 4, the nested gene set that we find from the *m2* intervals in the *ICA-3* system consists of *glpF* (b3927) - a glycerol MIP channel facilitator which allows the facilitated

diffusion of glycerol into the cell - in IC1-IC2; *smf* (b3285), and *ivbL* (b3672) - an operon leader peptide - in IC1-IC3; *glnH* (b0811), *nuoH* (b2282) - a subunit of NADH dehydrogenase I, *glpD* (b3426) - a subunit of glycerol 3-phosphate dehydrogenase (a homodimeric enzyme), and *treC* (b4239) - trehalose-6-phosphate hydrolase - in IC2-IC3.

Overall, *replicate 3* has the highest interference, followed by *replicate 1*; *replicate 2* is instead the sparsest one, and acts as the sparsity bound driver in the nesting process. A sub-group of interference genes refers to catabolite repression or activation mediated by the cyclic AMP receptor protein (*Crp*)[18].

It is of interest to check what is discarded from replicate 3, which is highly sparsified; *ndh* (b1109) and *yjiY* (b4354, a putative carbon starvation protein) in IC1-IC2; *yadN* (b0141, a putative fimbrial-like protein), *gltA* (b0720, a citrate synthase monomer), *mukB* (b0924, cell division protein involved in chromosome partitioning), again *ndh*, then *ymfD* (b1137), *hisL* (b2018, his operon leader peptide), *glpF* (b3927), *yjfO* (b4189, a conserved protein) in IC1-IC3; again *ndh* (b1109), *stfE* (b1157, a putative tail fiber protein), *glpK* (b3926, a subunit of glycerol kinase), *groL* (b4143, chaperone Hsp60), again *yjiY* (b4354) in IC2-IC3.

The following genes result instead selected in only two over three replicates: *ymfD* (b1137) in replicates 1 and 2, and *ndh* (b1109) in replicates 1 and 3, for IC1-IC2; *yjfO* (b4189) in replicates 2 and 3, *recN* (b2616, protein used in recombination and DNA repair) and *recX*, (b2698, inhibitor of *recA*, which is responsible for DNA strand exchange and renaturation) in replicates 1 and 2, *hisL* (b2018) in replicates 1 and 3 for IC1-IC3; (b3914, no information is available) in replicates 1 and 2, *stfE* (b1157) in replicates 1 and 3, *groL* (b4143) and *yjiY* (b4354) in replicates 2 and 3 for IC2-IC3.

In particular, the most stringent sparsity bound does not allow to consider both *recN* (b2616) and *recX* (b2698) that are detected in the doublet IC1-IC3 from the first two replicates, while they are assigned by *replicate 3* only to IC1, i.e. the most important component. Thus, due in this case to replicate 3, nesting is highly selective toward some truly relevant interference genes.

5.2.3 System with four components

In Table 5, the final nested gene set delivered by *ICA-4* is reported without interference of IC4 with any other component; in this case we cannot reach a bound, i.e. the selected genes are completely different from one replicate to another such that no intersection exists. Noise from an extra component should be clearly suspected to play a major role in these interference regions.

The relevant nested genes are *yjiY* (b4354) in IC1-IC2, *hisL* (b2018), *recN* (b2616), *smf* (b3285) and again *yjiY* (b4354) in IC1-IC3; *glnP* (b0810), *glnH* (b0811), *clpB* (b2592, clpB chaperone), *treC* (b4239) and again *yjiY* (b4354) in IC2-IC3. Overall, *replicate 3* is the one with the highest interference degree, followed by *replicate 1*. Both are more sparsified than *replicate 2*, which acts as the sparsity driver for the nesting process.

With genes selected from two over the three replicates, the interference region would include more genes, as before with the other ICA systems. By looking at the interference between the (presumably less noisy) first three components, from replicate 1 more genes

[18]Treatment of the cell with arabinose to induce the over-expression of genes on the plasmid might be a reason for *Crp* activation.

than from replicates 2 and 3 have been discarded.

In particular, *ndh* (b1109) in IC1-IC2 from replicates 1 and 2, *uraA* (b2497, UraA uracil NCS2 transporter) in IC1-IC3 from replicates 1 and 3 in IC1-IC3, and *glpD* (b3426) from replicates 1 and 2, *stfE* (b1157) from replicates 1 and 3. Then, *nuoH* (b2282) and *groL* (b4143) have been discarded from replicates 2 and 3 in IC2-IC3.

The interesting aspect addressed before with the *ICA-3* system, is present here too; both *recN* (b2616) and *recX* (b2698) appear in IC1-IC3, this time together only in replicate 2, otherwise split (*recN* in replicates 1 and 3). The former gene is thus selected in the nested gene set, and suggests a certain increased redundancy (i.e. three replicates select the same *SOS*-related gene) in *ICA-4* compared to *ICA-3*.

5.2.4 Comparisons

In order to see how the interference depends on each ICA system, we compare every doublet across replicates. First, IC1-IC2 is similar in all these systems as far as concerns replicate 1. When comparing the systems with more than two components, for both IC1-IC3 and IC2-IC3 each replicate presents a very similar selection of genes.

Thus, choosing a system with more than two components should be safer with regard to possible loss of informative genes; however, the fourth component is not contributing to nested interference as no sparsity bounds can be found in our data.

In summary, the aggregation via sieving of parallel feature selectors applied to our replicate data sets can be considered a useful data mining strategy.

In particular, it has revealed that:

- More dimensions are endowed with increased interference, but with more detrimental impact of noise;
- With more than two components, replicate 2 and replicate 3 appear to be the sparsity bound drivers most of the times, thus leading to the nested gene sets. In particular, replicate 2 more often than replicate 3 shows the lowest interference degree.

Thus, by discounting noise and unless any kind of prior information suggests that a very conservatively decomposed system should be adopted, we consider that more than minimal dimensionality (from only two components) may be in this application a better choice in order to account for more quality features.

Then, we stress the role of two over three replicates, as one of them appears less informative (*replicate 1*). Naturally enough, owing to a strict dependence of the data on the specific experimental outcomes, a generalization of these findings is not possible. However, an interesting consideration is that the aggregation and sieving methodological steps shed some light over useful aspects related to experimental design and data mining strategies.

5.3 Pattern Recognition

Figure 6 shows pre-nesting and post-nesting interference patterns computed at *m2* intervals for the component doublets across replicates[19]. The coordinate system of these plots consist

[19]For reasons of space, only the case of *ICA-4* is displayed.

of the sorted gene labels' identifiers in the ordinate, and the size of the gene interference group in the abscissa.

This visualization aims to emphasize the pattern similarity/dissimilarity of the selected gene groups. Given a certain component doublet, two extreme cases are: **A**) when an interference pattern repeats itself, i.e. for each replicate the same genes overlap across the two components; **B**) when instead completely different patterns reflect overlapping genes that change at each replicate.

In other terms, pattern dissimilarity suggests that there is replicate-dependent interference; therefore, relying on feature learning based on an average of information from individual replicates represents a risk for detecting the underlying gene dependence map. Vice versa, pattern similarity basically implies a more homogeneous interference that makes feature learning replicate-independent and averaging less harmful.

By looking at the plots, we can see that the absence of sparsity bounds, whenever IC4 is involved, is particularly visible in both IC2-IC4 and IC3-IC4 post-nesting patterns, but also in IC1-IC4 the sparsification produces dissimilarity. This is in contrast with what is observed for instance in IC1-IC3 or IC2-IC3, where the post-nesting patterns from replicate 1 seem redundant compared to those from the other replicates.

6 Concluding remarks

The probability of identifying differentially expressed genes depends on several factors, and especially on the true expression value (due to biological variation), the magnitude of noise fluctuations (from various sources), and the number of experimental replicates (aiming to increase the homogeneity in variability). It is about this latter aspect that we have conducted our work.

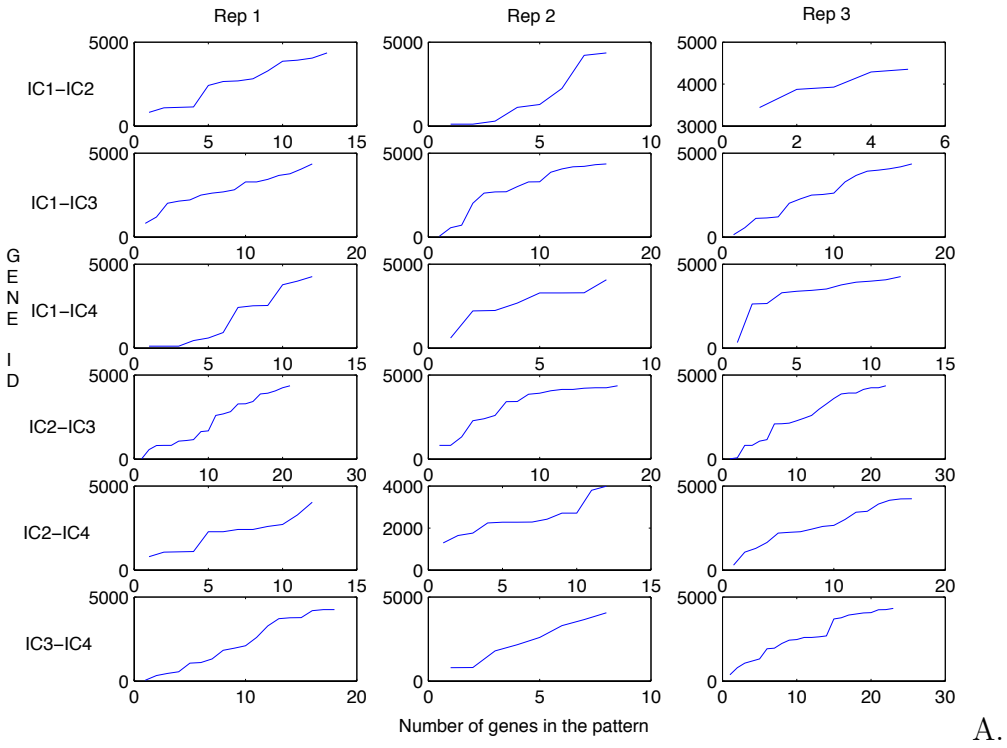
Clustering and dimensional reduction techniques can perform accurate gene feature selection, but understanding the nature of the interactions between genes is generally a very hard task for unsupervised learning techniques. When a feature learner such as ICA is employed in genomics, one aims to perform dimensionality reduction and select genes.

However, it is also important to explore the signal interference effects, which in this context consist of overlaps between features. Consequently, there is interest in examining genes that belong to more than one estimated component at the same time, as a result of the presence of noise and errors, or because reflecting the gene-gene interactions occurring between pathways or involving several biological functions and processes.

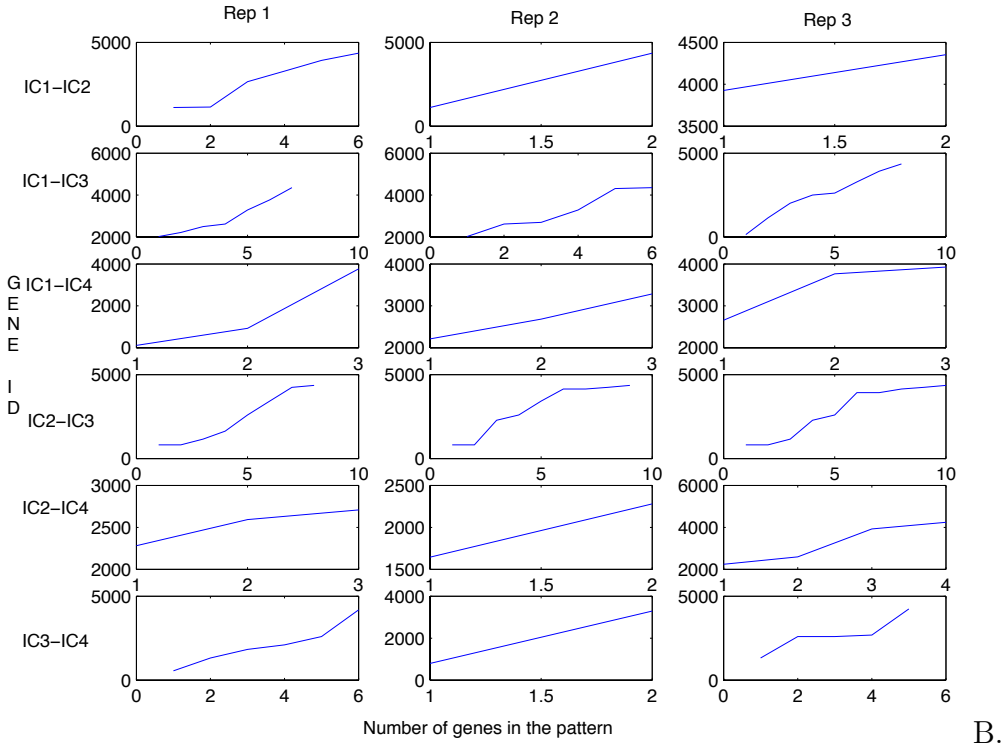
Elucidating these relationships might lead to better separation between gene co-expression and co-regulation dynamics. By looking at each individual replicate, we have a chance to use more information than that usually provided by an average of all of them. Also, we can measure the ICA's performance in parallel fashion, check the consistency of its results across replicates, and try to infer the various possible sources of interference.

The task here proposed is to reduce both redundancies and discrepancies, and then implement a simple heuristic nesting or sieving procedure so as to define a minimal informative core. The method may also be considered a search for sparsity bounds yielding a reference gene set to which the contributions of each replicate may be compared.

The outcomes of our application suggest that a more than minimal dimensionality (i.e.



A.



B.

Figure 6: *Pre-nesting* (A) and **post-nesting** (B) interference patterns for the *ICA-4* system.

a three-component ICA-based genome decomposition) performs better than smaller (too conservative in terms of components) or bigger (too redundant in terms of interference) systems.

Even if it is hard to generalize our results, we believe it is worth the attempt of questioning the informativeness of each replicate data, because a more or less marginal role can be assigned to a replicate compared to the others. With this regard, simply averaging the information across replicates may not represent the most accurate solution.

Conversely, the key aspect is whether ICA aggregation and feature stabilization via sieving can be used for eliminating redundancy, which brings advantages from a biological standpoint. It has been then shown that shrinking the interference regions through the nested gene set may be considered another useful reason that justifies a sparser treatment of the available data.

Acknowledgments

The author thanks the Biomedical Engineering department of Boston University for the experimental data. Financial support from *Sardegna Ricerche Tech Park* is gratefully acknowledged.

References

[1] Alter, O., Brown, P.O. and Botstein, D., Singular value decomposition for genome-wide expression data processing and modeling. *PNAS*, **97**, 10101-10106, 2000.

[2] Berger, J.A., Hautaniemi, S., Edgren, H., Monni, O., Mitra, S.K., Yli-Harja, O. and Astola, J., Identifying underlying factors in breast cancer using independent component analysis. *Proc. 2003 IEEE Int. Work. NNSP*, Toulose (FR) 81-90, 2003.

[3] Bofill, P. and Zibulevsky, M., Blind separation of more sources than mixtures using sparsity of their short-time fouries transform. *Proc. ICA*, 87-92, 2000.

[4] Breiman, L., Heuristics of instability and stabilization in model selection. *Ann. Statist.*, **24(6)**, 2350-2383, 1996.

[5] Breiman, L., Bagging Predictors. *Mach. Learn.*, **24(2)**, 123-140, 1996.

[6] Capobianco, E., Robustness vs Redundancy in Biological Systems. *Fluct. and Noise Lett.*, **5(3)**, 375-385, 2005.

[7] Capobianco, E., Mining Time-Dependent Gene Features. *J. Bioinform. Comput. Biol.*, **3(5)**, 1191-1205, 2005.

[8] Cardoso, J.F., Source separation using higher order moments. *Proc. ICASSP*, 2109-2112, 1989.

[9] Cardoso, J.F., Dependence, Correlation and Gaussianity in Independent Component Analysis. *J. Mach. Learn. Res.*, **4(7-8)**, 1177-1203, 2004.

[10] Cardoso, J.F. and Souloumiac, A., Blind beamforming for non-Gaussian signals. *IEE Proc. F.*, **140(6)**, 771-774, 1993.

[11] Chen, A. and Bickel, P.J., Efficient Independent Component Analysis (I). Tech. Rep. n. 634, Dept. of Statist., UC Berkeley (USA), 2003.

[12] Chen, A. and Bickel, P.J., Efficient Independent Component Analysis (II). Tech. Rep. n. 645, Dept. of Statist., UC Berkeley (USA), 2003.

[13] Chiappetta, P., Roubaud, M.C. and Torresani, B., Blind Source Separation and the Analysis of Microarray Data. *J. Comput. Biol.*, **11(6)**, 1090-109, 2004.

[14] Comon, P., Independent Component Analysis - a new concept? *Sig. Proces.*, **36(3)**, 287-314, 1994.

[15] Cover, T.M. and Thomas, J.A., *Elements of Information Theory* (New York: Wiley), 1981.

[16] http://ecocyc.org

[17] Friedman, J.H. and Tukey, J.H., A Projection Pursuit Algorithm for Exploratory Data Analysis. *IEEE Trans. Comput.*, **C-23**, 881-889, 1974.

[18] Gribonval, R., Vincent, E., Fevotte, C. and Benaroya, L., Proposals for performance measurement in source separation. *4th Int. Sympos. ICA - BSS*, Nara (JP), 2003.

[19] Holter, N., Maritan, A., Cieplak, M., Fedoroff, N. and Banavar, J., Dynamic modeling of gene expression data. *PNAS*, **98(4)**, 1693-1698, 2001.

[20] Hori, G., Inoue, M., Nishimura, S. and Nakahara, H., Blind gene classification. An application of a signal separation method. *Gen. Inform.*, **12**, 255-256, 2001.

[21] Hori, G., Inoue, M., Nishimura, S. and Nakahara, H., Blind gene classification. An ICA-based gene classification/clustering method. Tech. Rep. n. 02-5, Riken BSI/BSIS (JP), 2002.

[22] Huber, P.J., Projection Pursuit. *Ann. Statist.*, **13**, 435-475, 1985.

[23] Hyvarinen, A. and Oja, E., A fast fixed-point algorithm for Independent Component Analysis. *Neural Comput.*, **9(7)**, 1483-1492, 1997.

[24] Hyvarinen, A., Fast and robust fixed-point algorithms for Independent Component Analysis. *IEEE Trans. Neural Net.*, **10(3)**, 626-634, 1999.

[25] Jolliffe, I.Y., *Principal Component Analysis* (New York: Springer), 1986.

[26] Lee, S. and Batzoglou, S., Application of independent component analysis to microarrays. *Gen. Biol.*, **4:R76**, 2003.

[27] Liebermeister, W., Linear modes of gene expression determined by independent component analysis. *Bioinform.*, **18**, 51-60, 2002.

[28] Mika, S., Scholkopf, B., Smola, A., Muller, K.R., Scholz, M. and Ratsch, G., 1999, Kernel PCA and de-noising in feature spaces, In: Kearns, M.S., Solla, S.A. and Cohn, D.A. (Eds.), *Advances in NIPS* (Cambridge: MIT Press), **11**, 536-542, 1999.

[29] Rickard, S., Balan, R. and Rosca, J., Blind source separation based on space-time-frequency diversity. *ICA-BSS Conf.*, San Diego, CA, 2003.

[30] Scholkopf, B., Smola, A. and Muller, K.R., Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, **10(5)**, 1299-1319, 1998.

[31] Vrabie, V.D., Mars, J.I. and Lacoume, J.L., Modified Singular Value Decomposition by means of Independent Component Analysis. *Sig. Proces.*, **84**, 645-652, 2004.

[32] Zibulewsky, M. and Pearlmutter, B.A., Blind Source Separation by Sparse Decomposition in a Signal Dictionary. *Neural Comput.*, **13(4)**, 863-882, 2001.

Replicate 1										
IC1 vs IC2	*b*141	*b*0449	*b*0811	*b*1085	*b*1109	*b*1137	*b*1819	*b*1846	*b*2656	*b*3285
	*b*3856	*b*3927	*b*4053							
IC1 vs IC2	*b*0141	–	–	–	*b*1109	*b*1137	–	–	*b*2656	*b*3285
	–	**b3927**	–							
IC1 vs IC3	*b*0449	*b*0553	*b*0811	*b*1192	*b*1832	*b*2018	*b*2345	*b*2616	*b*2698	*b*3282
	*b*3285	*b*3441	*b*3521	*b*3672	*b*3766	*b*4053	*b*4060			
IC1 vs IC3	–	–	–	–	–	*b*2018	–	*b*2616	*b*2698	–
	b3285	–	–	**b3672**	*b*3766	–	–			
IC2 vs IC3	*b*0014	*b*0449	*b*0564	*b*0809	*b*0810	*b*0811	*b*1068	*b*1100	*b*1117	*b*1157
	*b*1633	*b*2239	*b*2282	*b*2685	*b*2818	*b*3285	*b*3426	*b*3427	*b*3870	*b*3914
	*b*4053	*b*4239								
IC2 vs IC3	–	–	–	–	–	**b0811**	–	–	–	*b*1157
	*b*1633	–	**b2282**	–	–	–	**b3426**	–	–	*b*3914
	–	**b4239**								
Replicate 2										
IC1 vs IC2	*b*0296	*b*0449	*b*1137	*b*1262	*b*1285	*b*3856	*b*3927	*b*4354		
IC1 vs IC2	–	–	*b*1137	–	–	–	**b3927**	–		
IC1 vs IC3	*b*0047	*b*0796	*b*2277	*b*2616	*b*2685	*b*2698	*b*3005	*b*3285	*b*3672	*b*3856
	*b*4189	*b*4214	*b*4311	*b*4354						
IC1 vs IC3	–	–	–	*b*2616	–	*b*2698	–	**b3285**	**b3672**	–
	*b*4189	–	–	–						
IC2 vs IC3	*b*0811	*b*1321	*b*2282	*b*2393	*b*2592	*b*3409	*b*3426	*b*3856	*b*3914	*b*4062
	*b*4143	*b*4214	*b*4239	*b*4240	*b*4354					
IC2 vs IC3	**b0811**	–	**b2282**	–	–	–	**b3426**	–	*b*3914	–
	*b*4143	–	**b4239**	–	*b*4354					
Replicate 3										
IC1 vs IC2	*b*1109	*b*3441	*b*3927	*b*4354						
IC1 vs IC2	*b*1109	–	**b3927**	*b*4354						
IC1 vs IC3	*b*0141	*b*0553	*b*0720	*b*0924	*b*1109	*b*1137	*b*1200	*b*2018	*b*2278	*b*2497
	*b*2535	*b*3280	*b*3285	*b*3672	*b*3694	*b*3766	*b*3927	*b*3932	*b*4068	*b*4189
	*b*4354									
IC1 vs IC3	*b*0141	–	*b*0720	*b*0924	*b*1109	*b*1137	–	*b*2018	–	–
	–	–	**b3285**	**b3672**	–	–	*b*3927	–	–	*b*4189
IC2 vs IC3	*b*0014	*b*0796	*b*0809	*b*0810	*b*0811	*b*1068	*b*1109	*b*1157	*b*2091	*b*2239
	*b*2282	*b*2393	*b*2592	*b*2964	*b*3281	*b*3426	*b*3603	*b*3870	*b*3926	*b*3927
	*b*4143	*b*4239	*b*4240	*b*4354						
IC2 vs IC3	–	–	–	–	**b0811**	–	*b*1109	*b*1157	–	–
	b2282	–	–	–	–	**b3426**	–	–	*b*3926	*b*3927
	*b*4143	**b4239**	–	*b*4354						

Table 4: Interference (*pre-nesting* and **post-nesting**) for *ICA-3*. The numbers indicate gene identifiers.

Replicate 1										
IC1 vs IC2	*b*0811	*b*1085	*b*1109	*b*1137	*b*2419	*b*2656	*b*2685	*b*2818	*b*3285	*b*3856
	*b*3927	*b*4053	*b*4354							
IC1 vs IC2	–	–	*b*1109	*b*1137	–	*b*2656	–	–	*b*3285	–
	*b*3927	–	**b4354**							
IC1 vs IC3	*b*0811	*b*1192	*b*2018	*b*2143	*b*2211	*b*2497	*b*2616	*b*2685	*b*2818	*b*3282
	*b*3285	*b*3441	*b*3672	*b*3766	*b*4053	*b*4354				
IC1 vs IC3	–	–	**b2018**	–	*b*2211	*b*2497	**b2616**	–	–	–
	b3285	–	–	*b*3766	–	**b4354**				
IC2 vs IC3	*b*0014	*b*0564	*b*0809	*b*0810	*b*0811	*b*1068	*b*1100	*b*1157	*b*1633	*b*1675
	*b*2592	*b*2685	*b*2818	*b*3281	*b*3285	*b*3426	*b*3870	*b*3914	*b*4053	*b*4239
	*b*4354									
IC2 vs IC3	–	–	–	**b0810**	**b0811**	–	–	*b*1157	*b*1633	–
	b2592	–	–	–	–	*b*3426	–	–	–	**b4239**
	b4354									
Replicate 2										
IC1 vs IC2	*b*0112	*b*0115	*b*0296	*b*1109	*b*1285	*b*2240	*b*4214	*b*4354		
IC1 vs IC2	–	–	–	*b*1109	–	–	–	**b4354**		
IC1 vs IC3	*b*0047	*b*0557	*b*706	*b*2018	*b*2616	*b*2685	*b*2698	*b*3005	*b*3285	*b*3292
	*b*3856	*b*4060	*b*4189	*b*4214	*b*4311	*b*4354				
IC1 vs IC3	–	–	–	**b2018**	**b2616**	–	*b*2698	–	**b3285**	–
	–	–	–	–	*b*4311	**b4354**				
IC2 vs IC3	*b*0810	*b*0811	*b*1321	*b*2282	*b*2393	*b*2592	*b*3409	*b*3426	*b*3856	*b*3914
	*b*4062	*b*4142	*b*4143	*b*4214	*b*4239	*b*4240	*b*4354			
IC2 vs IC3	**b0810**	**b0811**	–	*b*2282	–	**b2592**	–	*b*3426	–	–
	–	*b*4142	*b*4143	–	**b4239**	–	**b4354**			
Replicate 3										
IC1 vs IC2	*b*3441	*b*3873	*b*3927	*b*4290	*b*4354					
IC1 vs IC2	–	–	*b*3927	–	**b4354**					
IC1 vs IC3	*b*0141	*b*0553	*b*1109	*b*1137	*b*1200	*b*2018	*b*2278	*b*2497	*b*2535	*b*2616
	*b*3285	*b*3672	*b*3927	*b*3988	*b*4068	*b*4189	*b*4354			
IC1 vs IC3	*b*0141	–	–	*b*1137	–	**b2018**	–	*b*2497	–	**b2616**
	b3285	–	*b*3927	–	–	–	**b4354**			
IC2 vs IC3	*b*0014	*b*0075	*b*0810	*b*0811	*b*1068	*b*1157	*b*2091	*b*2096	*b*2133	*b*2282
	*b*2436	*b*2592	*b*2964	*b*3281	*b*3603	*b*3870	*b*3926	*b*3927	*b*4143	*b*4239
	*b*4240	*b*4354								
IC2 vs IC3	–	–	**b0810**	**b0811**	–	*b*1157	–	–	–	*b*2282
	–	**b2592**	–	–	–	–	*b*3926	*b*3927	*b*4143	**b4239**
	–	**b4354**								

Table 5: Interference (*pre-nesting* and **post-nesting**) for *ICA-4*. The numbers indicate gene identifiers.